Finding a Home for Your Science

April 29, 2016

Like fog lifting from a lake, the haze that historically has shrouded the process of conducting psychological science is evaporating. Psychological scientists are recognizing that documenting the results of a study in a published paper isn't always enough — rather, for research to be as reproducible as possible, research practices and statistical analyses must be transparent. Likewise, both data and materials should be made public and available for other researchers to examine.

Creating open data is not as simple as it sounds, though. APS's Open Data badge (see sidebar, below) is not awarded to anyone who simply puts their data on a personal website or blog page or in a Dropbox folder. This is because these kinds of websites are not a reliable, persistent place for data to reside. Psychological scientists should use an approved data repository to store their data — one that guarantees longevity and quality archival. Fortunately, many services have sprung up to fill this need, and they all

have one thing in common: a commitment to protecting data from service shutdowns and hard-drive failures. With these services, “The dog ate my data!” is no longer a valid excuse. Best of all, many of them are free to use.

FYI on DOIs

The APS Open Practices guidelines state that an Open Data badge requires “a URL, DOI, or other permanent path for accessing the data in a public, open-access repository.” All readers are familiar with URLs (uniform resource locators), which specify web addresses. But the DOI — which stands for digital object identifier — is less widely known. These identifiers, which function somewhat like book ISBN numbers, link to URLs. The idea is that one DOI will always point to a specific object, like a data set or journal article, even if the location or characteristics of that object change. The DOI “10.1111/j.1467-9280.1990.tb00056.x” will always point to *Psychological Science*’s first article, an editorial written by Founding Editor William K. Estes, an APS William James Fellow, even if the website or publisher changes.

This behind-the-scenes magic results from the work of a number of organizations. DOIs are overseen and operated by the International DOI Foundation, which began assigning DOIs in 2000. A DOI factsheet details that DOIs have been assigned by more than 5,000 groups and that 120 million DOIs have been assigned as of this year. The International DOI Foundation oversees a group of DOI registration agencies that curate DOIs from different domains. Psychologists may recognize CrossRef, which covers journal articles, books, and other scholarly content, or DataCite, a newer group that focuses specifically on data. Other registration agencies focus mostly on products in Japan, Korea, and China.

These registration agencies accept the responsibility of ensuring that the link between a DOI and the information it connects to will remain stable — that the DOI always points to the right material. Section 6.5 of the DOI handbook reads, “Persistence of DOI information is a key aim of the DOI system, and is guaranteed by the DOI social infrastructure, policies, and agreements. In the event of any [registration agency] ceasing to maintain DOI information, for any reason, the records will be transferred to another [registration agency].”

The implication is that DOIs denote serious business — a commitment to making sure a link stands the test of time. When you see a DOI in print, you can be sure you can always find that piece of content. That’s why APS strongly encourages that DOIs be included for each academic paper in a reference list.

Other kinds of identifiers include ARKs (archival resource keys) and PURLs (persistent URLs). Despite differences, all of these identifiers ensure links go to the same place every time.

The Online Storage Lockers

A DOI is just a persistent link to a particular resource. What are some different data repositories that you can use to store your work?

The Open Science Framework (OSF) is both a project-management system and a repository for data and

materials. OSF users can upload all sorts of files, and when they are ready to share those materials publicly, they can "register" the project. This creates a permanent, public record. The OSF is free and has no storage limit. Once a project is registered, users can click a button to generate a DOI and ARK for that project and then share or cite those data or materials.

Figshare is a file-sharing site and data repository that provides users with unlimited public storage space and upload thresholds for files up to 5 gigabytes (GB) in size (users can store up to 20 GB of private files, too). Accounts are free, and researchers can generate DOIs for their projects on this site. Figshare has a streamlined project-management component, which users may find to be simpler, but not as extensive.

Dryad is another repository for data or research code, especially for research involving the life sciences. Unlike OSF and Figshare, Dryad accepts submissions only for published articles, which are integrated with articles by Dryad's curation staff. Submitting data to Dryad costs $120, although submissions in conjunction with some journals (such as *Scientific Data*) are free. DOIs are also included with Dryad submissions.

Other options for data sharing abound. Many of them emanate from colleges and universities. These include Merritt, organized by the University of California Curation Center; the Dataverse Project, a collaboration from Harvard University; and the Inter-university Consortium for Political and Social Research, a product of an institute at the University of Michigan.

Psychological scientists looking for other options best tailored to their data set can visit the Registry of Research Data Repositories (r3data) to find more. Visitors to this site will find specialized data sets that they can use in their own research, too, such as Washington University in St. Louis's English Lexicon Project, which contains subject data for more than 40,000 words. Another repository, called Databrary, focuses on sharing and reuse of research videos. You can read more about Databrary in the March issue of the *Observer*.

What's on the Menu?

Once you've decided on a data repository, it's time to determine what kind of data to make available. This can be another tricky choice, because you can choose from a continuum of ways to submit data:

Raw Data — the output taken straight from the source (comma-separated-value or Excel files, online server output, etc.). Raw, unedited data are the purest kind of data to include, and are also the most transparent. They can, however, be difficult for other researchers to interpret or use, so careful annotation of the data is key.

Processed data — data that have been edited, "scrubbed," or otherwise organized by the researcher. The degree of processing, and choices made during this step, are at a researcher's discretion. Data can go through light processing — say, sorted into columns with added headers — or heavier processing; for instance, you could eliminate faulty trials or delete data from participants who did not follow instructions. These data may be easier for other scientists to interpret and use, but you may need to indicate the extent of processing that occurred during your preparation of the file.

Aggregated data — data that have been collapsed or combined in some way. This could involve combining trial-level data for individuals by averaging across trials and coming up with one score per subject, for example. Aggregated data can be easiest to interpret and reuse, perhaps even by the general public. But it's impossible to learn more about the underlying data that went into the aggregate, which means that other scientists are less able to evaluate those data.

Of course, researchers can post all three kinds of data on a repository, too. Receipt of APS's Open Data badge requires that researchers report sufficient information for an independent researcher to reproduce the reported results. The definition of "sufficient information" may vary on a paper-by-paper basis.

Post the First Time for Forever

Repositories like Figshare make varying types of commitments to keeping your data persistent and available.

"Figshare itself is built on Amazon Web Services … our infrastructure is distributed globally, allowing clients to decide where they wish to have their data stored — prioritizing resilience and backup or destination as they choose. We have more than 20 servers in our infrastructure. We've ensured all public end-user Figshare content is further archived and preserved" using the Digital Preservation Network, says Dan Valen, a Product Manager at Figshare.

Similarly, the Center for Open Science notes that OSF "is backed by a $250,000 preservation fund that will provide for persistence of your data, even if the Center for Open Science runs out of funding." It also offers details describing the technologies that are used for backup, such as Amazon's Glacier platform, a service designed to back up data that don't need to be retrieved regularly (called "cold data").

Repositories may take other steps or use other methods to keep content backed up. For instance, some repositories take advantage of an archive called LOCKSS, which stands for "Lots of Copies Keep Stuff Safe" — the same kind of archive that helps back up journals and books. The idea behind LOCKSS is that when scholarly content is put online, it is copied and preserved by other trusted backup systems in case something should happen to the original. This distributed system avoids single points of failure.

(In fact, the slogan "lots of copies keep stuff safe" is a valuable one to keep in mind when considering your own data backup-storage plan. And that applies to more than just experimental data.)

Ethical Concerns

If you are indeed planning to submit your data to a repository, you may face some associated ethical hurdles. For instance, subjects may not want their data posted online, even if they are in no way identified. Or a researcher may decide to make data available online only after data collection has occurred and subjects are not available to consent to sharing their data. In these cases, you may need to work with institutional review boards to ensure that it is appropriate to make data public. Thus, it's best to anticipate making data open in advance, prior to obtaining ethical approval.

Indeed, ethics institutions are paying heightened attention to the value and risks of data sharing. A notice of proposed rulemaking for the Common Rule, a federal document that guides human-subjects research regulations, states:

“People share information about themselves with large numbers of people with the click of a button, and this trend of rapid and widespread sharing is only likely to grow. The increase in concern about unauthorized and inadvertent information disclosure, in combination with newer research techniques that increase the volume and nature of identifiable data, suggest the need for the Common Rule to more explicitly address data security and privacy protection.”

Oops, I Undid It Again

What happens when a researcher, after putting data online, wants to change or delete them? There is usually a way to remove data on even the most secure of sites. For instance, data registered on the OSF can be retracted, similar to how an article might get retracted from a journal.

“Retracting a registration will remove its content from the OSF, but leave basic metadata behind. The title of a retracted registration and its contributor list will remain, as will justification or explanation of the retraction, should you wish to provide it. Retracted registrations will be marked with a retracted tag,” the OSF website reads.

Thus, researchers can remove even registered data from the OSF, but it will be obvious to viewers that the data were retracted. Is this always a negative outcome? In some cases, retracting public data might be appropriate: If the data contained unapproved identifying information, or private health information that was not meant to be shared, the potential harm caused by leaving the files online outweighs the trouble caused by taking them down.

Likewise, Figshare reports, “If you accidentally make research publicly available or if there are ethical or legal reasons why this research should not be open, we adhere to the standard COPE guidelines in the same way as traditional academic publishers.” COPE, the Committee on Publication Ethics, develops best practice guidelines for editors and publishers regarding ethical matters.

Safely Ejecting This Story

Making data public may seem straightforward, but the process can be a tricky one to navigate. Authors must choose a repository, ensure ethical approval, upload files, and then cite those data correctly. But regardless of the exact method you use to make data available, the contribution you make to scholarship and reproducibility benefits all. œ

Further Reading

Gewin, V. (2016). Data sharing: An open mind to open data. *Nature*, *529*, 117–119. doi:10.1038/nj7584-117a

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A.

(2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science, 3*, 150547. doi:10.1098/rsos.150547

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE, 2*: e308. doi:10.1371/journal.pone.0000308

Wicherts, J. (2013). Science revolves around the data. *Journal of Open Psychology Data, 1*, 1–4. doi:10.5334/jopd.e1